(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 January 2007 (18.01.2007)

PCT

(10) International Publication Number
WO 2007/008492 A2

(51) International Patent Classification:
G06F 17/27 (2006.01)

(21) International Application Number:
PCT/US2006/026012

(22) International Filing Date: 30 June 2006 (30.06.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/177,136 8 July 2005 (08.07.2005) US

(71) Applicant *(for all designated States except US)*: **MICROSOFT CORPORATION** [US/US]; One Microsoft Way, Redmond, Washington 98052-6399 (US).

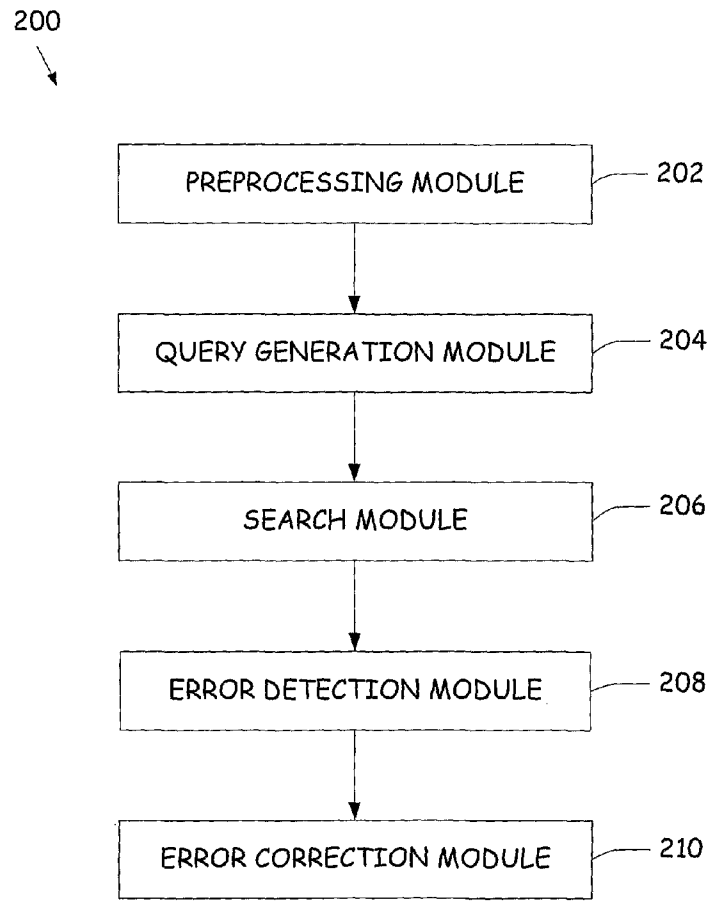
(72) Inventors: **HON, Hsiao-Wuen**; One Microsoft Way, Redmond, Washington 98052-6399 (US). **GAO, Jianfeng**; One Microsoft Way, Redmond, Washington 98052-6399 (US). **ZHOU, Ming**; One Microsoft Way, Redmond, Washington 98052-6399 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: PROCESSING COLLOCATION MISTAKES IN DOCUMENTS



(57) Abstract: A sentence is accessed and at least one query is generated based on the sentence. At least one query can be compared to text within a collection of documents, for example using a web search engine. Collocation errors in the sentence can be detected and/or corrected based on the comparison of the at least one query and the text within the collection of documents.

WO 2007/008492 A2

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

PROCESSING COLLOCATION MISTAKES IN DOCUMENTS

BACKGROUND

The discussion below is merely provided for general background information and is not intended to be used as an aid in determining the scope of the claimed subject matter.

With an ever increasing global economy, and with the rapid development of the Internet, people all over the world are becoming increasingly familiar with writing in a language which is not their native language. Unfortunately, for some societies that possess significantly different cultures and writing styles, the ability to write in some non-native languages is an ever-present barrier. When writing in a non-native language (for example English), language usage mistakes are frequently made by non-native speakers (for example, people who speak Chinese, Japanese, Korean or other non-English languages). These kind of mistakes can include both grammatical mistakes and improper usage of collocations such as verb-object, adjective-noun, adverb-verb, etc.

Many people have the ability to write in a non-native language using proper grammar, but they still may struggle with mistakes in collocations between two words. Still others struggle with both grammar and other mistakes such as collocations between two words. While spell checking and grammar checking programs are useful in correcting grammatical mistakes, detection and/or correction of mistakes in collocations between two words can be difficult, particularly since these mistakes can be otherwise grammatically correct. Therefore, grammar

checkers typically provide very little assistance, if any, in detecting mistakes relating to the collocation between words. English is used as an example of the non-native language in the following discussion, but these problems persist across other language boundaries.

For example, consider the following sentences that contain collocation mistakes which cause the sentences to not be native-like English, even if otherwise grammatically correct.

1. Open the light.

2. Everybody hates the crowded traffic on weekends.

3. This is a check of US$ 500.

4. I congratulate you for your success.

The native-like English versions of these sentences should be like:

1. Turn on the light.

2. Everybody hates the heavy traffic on weekends.

3. This is a check for US$ 500.

4. I congratulate you on your success.

As an example of the barriers faced by non-native English speaking peoples, consider the plight of the Chinese user. By culture, background and thinking habits,

Chinese people often produce English sentences which may be grammatical, but not natural. For example, Chinese people tend to directly translate subjects in Chinese into subjects in English, and do the same with objects and verbs. When writing in English, Chinese people often experience difficulty in deciding the collocations between verbs and prepositions, adjectives and nouns, verbs and nouns, etc. Moreover, in specific domains like the business domain, special writing skills and styles are needed.

Common dictionaries are mainly used by non-native speakers for the purpose of reading (a kind of decoding process), but these dictionaries do not provide enough support for writing (a kind of encoding process). They only provide the explanation of a single word, and they typically do not provide sufficient information to explain relevant phrases and collocations. Moreover, there is no easy way to get this kind of information from dictionaries, even if some of the information is provided in the dictionaries. On the other hand, current widely used grammar checking tools have some limited ability in detecting apt-to-make grammatical mistakes, but are not able to detect the collocation mistakes.

SUMMARY

This Summary is provided to introduce some concepts in a simplified form that are further described below in the Detailed Description. This Summary is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used as an aid in determining the scope of the claimed subject matter.

A sentence is accessed and at least one query is generated based on the sentence. At least one query can be compared to text within a collection of documents, for example using a web search engine. Collocation errors in the sentence can be detected and/or corrected based on the comparison of the at least one query and the text within the collection of documents.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of a general computing environment.

FIG. 2 is a flow diagram of a system for detecting and correcting collocation errors.

FIG. 3 is a flow diagram of a method for detecting and correcting collocation errors.

FIG. 4 is a block diagram of a query generation module.

FIG. 5 is a flow diagram of a method for detecting collocation errors.

FIG. 6 is a flow diagram of a method for presenting candidate collocation corrections.

DETAILED DESCRIPTION

FIG. 1 illustrates an example of a suitable computing system environment 100 on which the invention may be implemented. The computing system environment 100 is only one example of a suitable computing environment and is not intended to suggest any limitation as to the scope of use or functionality of the invention. Neither should the computing environment 100 be interpreted as having any dependency or requirement relating to any one

or combination of components illustrated in the exemplary operating environment 100.

The invention is operational with numerous other general purpose or special purpose computing system environments or configurations. Examples of well-known computing systems, environments, and/or configurations that may be suitable for use with the invention include, but are not limited to, personal computers, server computers, hand-held or laptop devices, multiprocessor systems, microprocessor-based systems, set top boxes, programmable consumer electronics, network PCs, minicomputers, mainframe computers, telephony systems, distributed computing environments that include any of the above systems or devices, and the like.

The invention may be described in the general context of computer-executable instructions, such as program modules, being executed by a computer. Generally, program modules include routines, programs, objects, components, data structures, etc. that perform particular tasks or implement particular abstract data types. The invention may also be practiced in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in both local and remote computer storage media including memory storage devices. Tasks performed by the programs and modules are described below and with the aid of figures. Those skilled in the art can implement the description and figures as processor executable instructions, which can be written on any form of a computer readable medium.

With reference to FIG. 1, an exemplary system for implementing the invention includes a general-purpose computing device in the form of a computer 110. Components of computer 110 may include, but are not limited to, a processing unit 120, a system memory 130, and a system bus 121 that couples various system components including the system memory to the processing unit 120. The system bus 121 may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. By way of example, and not limitation, such architectures include Industry Standard Architecture (ISA) bus, Micro Channel Architecture (MCA) bus, Enhanced ISA (EISA) bus, Video Electronics Standards Association (VESA) local bus, and Peripheral Component Interconnect (PCI) bus also known as Mezzanine bus.

Computer 110 typically includes a variety of computer readable media. Computer readable media can be any available media that can be accessed by computer 110 and includes both volatile and nonvolatile media, removable and non-removable media. By way of example, and not limitation, computer readable media may comprise computer storage media and communication media. Computer storage media includes both volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as computer readable instructions, data structures, program modules or other data. Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical disk storage, magnetic cassettes,

magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store the desired information and which can be accessed by computer 110. Communication media typically embodies computer readable instructions, data structures, program modules or other data in a modulated data signal such as a carrier wave or other transport mechanism and includes any information delivery media. The term "modulated data signal" means a signal that has one or more of its characteristics set or changed in such a manner as to encode information in the signal. By way of example, and not limitation, communication media includes wired media such as a wired network or direct-wired connection, and wireless media such as acoustic, RF, infrared and other wireless media. Combinations of any of the above should also be included within the scope of computer readable media.

The system memory 130 includes computer storage media in the form of volatile and/or nonvolatile memory such as read only memory (ROM) 131 and random access memory (RAM) 132. A basic input/output system 133 (BIOS), containing the basic routines that help to transfer information between elements within computer 110, such as during start-up, is typically stored in ROM 131. RAM 132 typically contains data and/or program modules that are immediately accessible to and/or presently being operated on by processing unit 120. By way of example, and not limitation, FIG. 1 illustrates operating system 134, application programs 135, other program modules 136, and program data 137.

The computer 110 may also include other removable/non-removable volatile/nonvolatile computer

storage media. By way of example only, FIG. 1 illustrates a hard disk drive 141 that reads from or writes to non-removable, nonvolatile magnetic media, a magnetic disk drive 151 that reads from or writes to a removable, nonvolatile magnetic disk 152, and an optical disk drive 155 that reads from or writes to a removable, nonvolatile optical disk 156 such as a CD ROM or other optical media. Other removable/non-removable, volatile/nonvolatile computer storage media that can be used in the exemplary operating environment include, but are not limited to, magnetic tape cassettes, flash memory cards, digital versatile disks, digital video tape, solid state RAM, solid state ROM, and the like. The hard disk drive 141 is typically connected to the system bus 121 through a non-removable memory interface such as interface 140, and magnetic disk drive 151 and optical disk drive 155 are typically connected to the system bus 121 by a removable memory interface, such as interface 150.

The drives and their associated computer storage media discussed above and illustrated in FIG. 1, provide storage of computer readable instructions, data structures, program modules and other data for the computer 110. In FIG. 1, for example, hard disk drive 141 is illustrated as storing operating system 144, application programs 145, other program modules 146, and program data 147. Note that these components can either be the same as or different from operating system 134, application programs 135, other program modules 136, and program data 137. Operating system 144, application programs 145, other program modules 146, and program data 147 are given different numbers here to illustrate that, at a minimum, they are

different copies.

A user may enter commands and information into the computer 110 through input devices such as a keyboard 162, a microphone 163, and a pointing device 161, such as a mouse, trackball or touch pad. Other input devices (not shown) may include a joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 120 through a user input interface 160 that is coupled to the system bus, but may be connected by other interface and bus structures, such as a parallel port, game port or a universal serial bus (USB). A monitor 191 or other type of display device is also connected to the system bus 121 via an interface, such as a video interface 190. In addition to the monitor, computers may also include other peripheral output devices such as speakers 197 and printer 196, which may be connected through an output peripheral interface 190.

The computer 110 may operate in a networked environment using logical connections to one or more remote computers, such as a remote computer 180. The remote computer 180 may be a personal computer, a hand-held device, a server, a router, a network PC, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer 110. The logical connections depicted in FIG. 1 include a local area network (LAN) 171 and a wide area network (WAN) 173, but may also include other networks. Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets and the Internet.

When used in a LAN networking environment, the

computer 110 is connected to the LAN 171 through a network interface or adapter 170. When used in a WAN networking environment, the computer 110 typically includes a modem 172 or other means for establishing communications over the WAN 173, such as the Internet. The modem 172, which may be internal or external, may be connected to the system bus 121 via the user input interface 160, or other appropriate mechanism. In a networked environment, program modules depicted relative to the computer 110, or portions thereof, may be stored in the remote memory storage device. By way of example, and not limitation, FIG. 1 illustrates remote application programs 185 as residing on remote computer 180. It will be appreciated that the network connections shown are exemplary and other means of establishing a communications link between the computers may be used.

FIG. 2 is a flow diagram of a system 200 for detecting and correcting collocation errors within text. There are many types of collocation errors. In one aspect of system 200, four types of collocation errors are detected. The collocation error types include:

1. verb-noun (VN, e.g. *learn/acquire knowledge),
2. preposition-noun (PN, e.g. *on/in the morning),
3. adjective-noun (AN, e.g. *social/socialist country), and
4. verb-adverb (VA, e.g. situations change *largely/greatly).

Preprocessing module 202 processes text to provide part of speech tagging and parsing of the text. Many

different types of parsers can be used to process the text. Below is an example sentence:

I have recognized this person for years.

Preprocessing module 202 tags this sentence and chunks the sentence as follows:

[NP I/PRP][VP have/VBP recognized/VBN][NP this/DT person/NN][PP for/IN][NP years. </s> /NNS]

Using the processed text, query generation module 204 constructs queries. In one example, four sets of queries are generated for each type of collocation error type identified above. For example, the collocation error types can be verb-noun, preposition-noun, adjective-noun and verb-adverb. The queries generated can include the full text of the sentence as well as a reduced portion of the sentence where auxiliaries are removed. Example reduced queries for the sentence above can include, "have recognized this person", "have recognized", "this person" and "recognized person".

The queries are submitted to a search module 206. In one embodiment, the search module can be a web based search engine such as MSN Search (search.msn.com), Google (www.google.com) and/or Yahoo! (www.yahoo.com). Since the web includes a vast amount of text, it can be an inexpensive resource to detect collocation errors. Error detection module 208 compares the queries generated by query generation module 204 to results obtained by search module 206. Error correction module 210 provides candidate corrections for the errors identified by error detection module 208.

FIG. 3 is a flow diagram of a method 220 that can be implemented in system 200 illustrated in FIG. 2. At step 222, a sentence is accessed. The sentence may include

text that has been input into a word processor, for example Microsoft Word® available from Microsoft Corporation of Redmond, Washington. At step 224, the sentence is parsed into chunks and parts of speech within the sentence are identified. Then, queries are generated based on the parse at step 226. At step 228, the queries are submitted to a search engine, for example MSN Search, Google and/or Yahoo!. Collocation errors in the sentence are detected at step 230 by comparing the queries and results from the search engine. After detecting errors, ranked candidates for alternatives to the collocation errors are presented to the user at step 232.

FIG. 4 is a block diagram of query generation module 204 of FIG. 2. Query generation module 204 accepts a parsed sentence 240, for example a parsed sentence received from preprocessing module 202. Based on parsed sentence 240, query generation module 204 generates sentence queries 242, chunk queries 244 and word queries 246. Given the types of potential collocation errors identified above, a *checking word* (i.e. a *word* that potentially causes a collocation error) is detected as *follows: verb in type VN, preposition in type PN,* adjective in type AN, and adverb in type VA. Depending on the type, query generation module 204 generates a different set of queries as follows:

1. Sentence queries 242: original sentence and reduced sentence (by removing auxiliaries pre-defined for each type), called *S-Query*,
2. Chunk queries 242: corresponding chunk pairs in the sentence, called *C-Query*, and
3. Word queries 246: corresponding headword pairs in the sentence, called *W-Query*.

Example queries for the sentence "I have recognized this person for years" for type VN detection are presented below, where ~ means that two adjacent words can be adjacent or one-word away from each other.

S-Query: ["I have recognized this person for years"]

S-Query: ["have recognized this person"]
C-Query: ["have recognized" ~ "this person"]
W-Query: ["recognized" ~ "person"]

Example rules for generating queries of each type are as follows.

- VN: S-Queries, one C-Query V~N, and one W-Query V_h~N_h (N_h denotes a headword of a corresponding noun chunk).
- PN: one C-Query of PN, which contains the preposition;
- AN: one C-Query of AN, which contains the AN pair; and
- VA: C-Queries that contain VA pairs and W-Queries that contains VA headwords.

FIG. 5 is a flow diagram of a method 250 of detecting errors in a sentence. Queries generated by query generation module 204 are submitted to the search module 206 at step 251. Search results obtained by the search module 206 are compared to the queries. In one example, the results include a summary of text for a document retrieved using a web search engine. At step 252, S-queries 242 from query generation module 204 are compared to results from the search module. Then, at step 254, a determination is made as to whether one or more of the S-queries 242 match the search module results. If one or more of the S-queries match the

search module results, it is determined that no collocation error exists at step 256.

However, if a match does not exist, method 250 proceeds to step 258, wherein C-queries 244 are compared to the search module results. At step 260, it is determined whether one or more of the C-queries closely match the search module results and if a score for the comparison is greater than a threshold. In one example, the score is computed by dividing a number of times the chunk of the C-query appears in the search results by the number of times words in the C-query co-occur in the search results. If the score is greater than the threshold, it is determined that no collocation error exists at step 256.

If the score is less than the threshold, method 250 proceeds to step 262, wherein W-Queries are compared to the search engine data. Step 264 determines whether there is a close match between the W-queries and the search engine data and whether a score for the comparison is greater than a threshold. If the score is greater than the threshold, it is determined that no collocation error exists at step 256. The score for the comparison can be similar to the C-Query comparison score. Thus, the W-Query comparison score can be calculated by dividing the number of times the W-Query occurs in the search results by the total number of times the pair of words in the W-Query co-occurs. If the score is less than the threshold, method 250 proceeds to step 266, wherein the user is notified of a potential collocation error.

FIG. 6 is a flow diagram of a method 270 for presenting potential corrected collocations to a user.

At step 272 a query template is generated. The query templates are generated based on a word that has been identified as an error (i.e. the checking word above includes a collocation error as determined by method 250 in FIG. 5). Query templates are derived from the input sentence after the checking word causing the collocation error has been replaced with a "+". In the sentence above, "recognized" has been identified as the checking word and thus query templates are developed based on the word. For example, the query templates of the sentence "I have recognized this person for years", for VN detection, are as follows, where + denotes any word.

S-QT: ["I have + this person for years"]

S-QT: ["I have + this person"]

S-QT: ["have + this person for years"]

S-QT: ["I have + this person"]

C-QT: ["+ this person for years"]

C-QT: ["+ this person"]

Example rules for generating query templates can be as follows.

- VN: S-QT, C-QT (where the verb has been replaced with +).
- PN: S-QT, C-QT (where the preposition has been replaced with +);
- AN: S-QT, C-QT (where the adjective has been replaced with +); and
- VA: S-QT, C-QT (where the adverb has been replaced with +).

At step 274, the query template is submitted to a search module, herein a search engine. At step 276, strings from the search engine results are retrieved. The strings can include summaries of text that have words of surrounding context. Strings that match the query templates, where the position of + can be any one word, are identified as strings *candidates*. Candidates that do not contain the collocation (which is formed by a word replacing + and another word in the string according to the collocation type) are removed at step 278. Remaining candidates are ranked according to a score based on a corresponding weight of query template that matched the string candidate. For example, the weight of the query template can be based on the number of words in the query template. The score for each candidate is calculated by taking a sum for weights across all summaries containing the candidates. The score for query templates (QT_s) that retrieve the candidate can be expressed by:

$$\text{Score}(candidate) = \sum_{QTs} \text{Weight}(QT)$$

A ranked list of candidates is then presented to a user at step 280. For example, a pop-up menu can be used to present the ranked list. A user can choose one of the selections from the list to correct the collocation error.

Although the subject matter has been described in language specific to structural features and/or methodological acts, it is to be understood that the subject matter defined in the appended claims is not necessarily limited to the specific features or acts described above. Rather, the specific features and acts

described above are disclosed as example forms of implementing the claims.

WHAT IS CLAIMED IS:

1. A method of processing information in a document, comprising:

accessing a sentence of text in the document;

generating at least one query based on the sentence;

comparing the at least one query to text within a collection of documents; and

detecting collocation errors based on the comparison of the at least one query and the text within the collection of documents.

2. The method of claim 1 and further comprising parsing the sentence to identify parts of speech contained therein and wherein generating the at least on query is based on the identified parts of speech.

3. The method of claim 1 and further comprising presenting candidates of alternatives to the detected collocation errors.

4. The method of claim 3 and further comprising ranking the candidates.

5. The method of claim 1 wherein a type of the collocation errors includes at least one of verb-noun, preposition-noun, adjective-noun and verb-adverb.

6. The method of claim 1 and further comprising generating a sentence query including a sentence based on the sentence of text, generating a chunk query including

chunks of the sentence of text and generating a word query including a head word pair of the sentence of text.

7. A method of processing information comprising:

accessing a sentence including a word that creates a collocation error;

generating at least one query including portions of the sentence without the word;

submitting the at least one query to a search module to obtain search results; and

identifying at least one candidate replacement word for the word in the sentence based on the search results.

8. The method of claim 7 and further comprising identifying a list of candidate replacement words and ranking the list.

9. The method of claim 7 wherein the search module is a web based search engine.

10. The method of claim 9 wherein the web based search engine includes at least one of MSN Search, Google, and Yahoo!.

11. The method of claim 7 wherein generating at least one query includes generating a sentence query including the sentence without the word.

12. The method of claim 7 wherein generating at least one query includes generating a chunk query including a chunk of the sentence without the word.

13. The method of claim 7 wherein generating at least one query includes generating a word query including a head word pair of the sentence without the word.

14. The method of claim 7 wherein generating at least one query includes generating a template of queries based on the sentence, chunks of the sentence and a headword pair of the sentence.

15. A method of processing information in a document, comprising:

accessing a sentence;

parsing the sentence to identify parts of speech contained therein;

generating a plurality of queries based on the identified parts of speech;

submitting the plurality of queries to a search module to obtain search results;

comparing the plurality of queries and the search results; and

detecting collocation errors in the sentence based on the comparison and the identified parts of speech.

16. The method of claim 15 wherein the plurality of queries include a sentence query including a sentence, a chunk query including a chunk of the sentence and a word query including a headword pair of the sentence.

17. The method of claim 15 wherein parsing the sentence includes identifying at least one of a verb-noun pair, a preposition-noun pair, an adjective-noun pair and a verb-adverb pair.

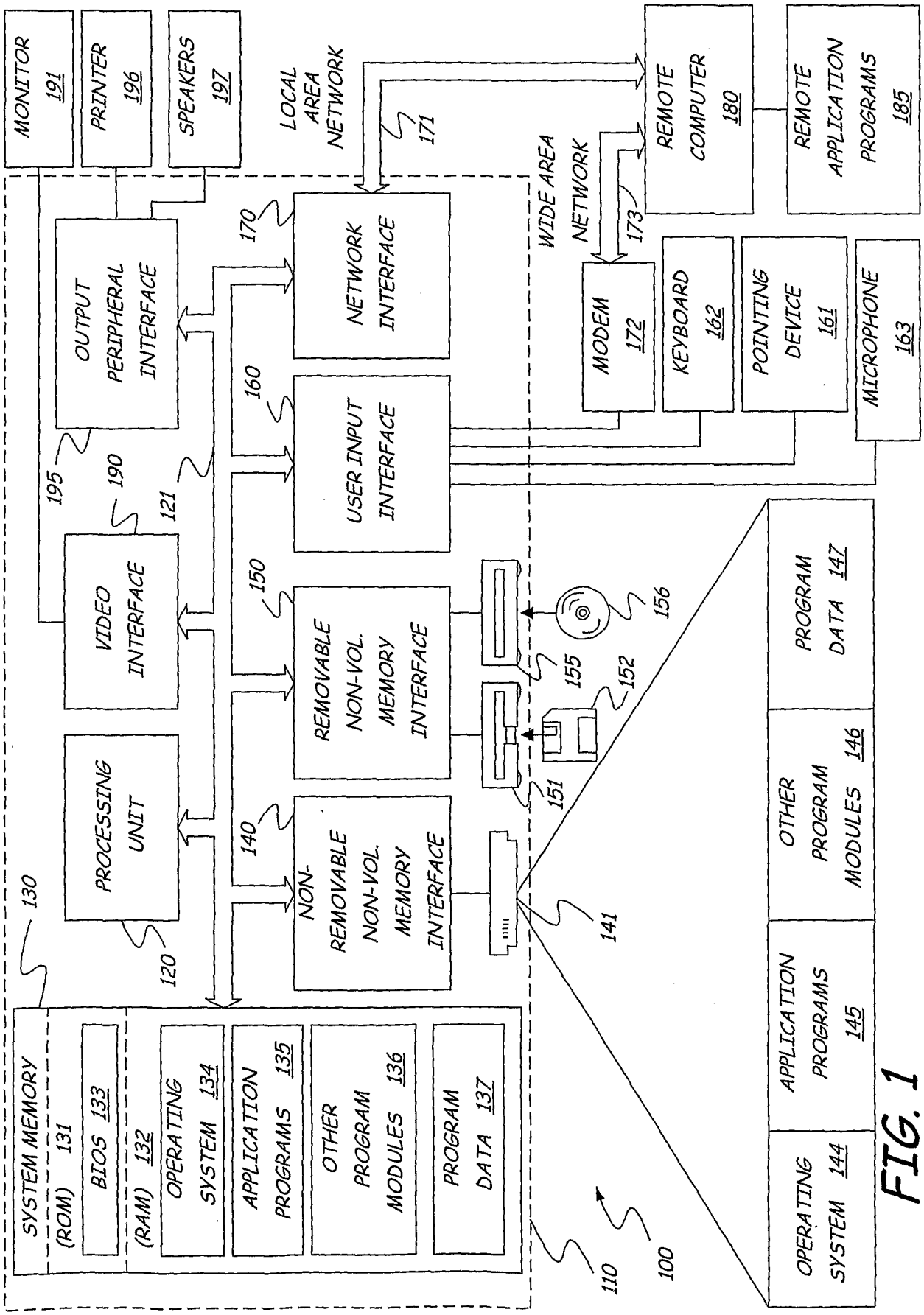

MONITOR 191
PRINTER 196
SPEAKERS 197
LOCAL AREA NETWORK
171
WIDE AREA NETWORK
REMOTE COMPUTER 180
REMOTE APPLICATION PROGRAMS 185
OUTPUT PERIPHERAL INTERFACE
NETWORK INTERFACE
170
MODEM 172
KEYBOARD 162
POINTING DEVICE 161
MICROPHONE 163
USER INPUT INTERFACE
160
195
190
121
VIDEO INTERFACE
PROCESSING UNIT
120
150
REMOVABLE NON-VOL. MEMORY INTERFACE
140
NON-REMOVABLE NON-VOL. MEMORY INTERFACE
155
156
152
151
130
SYSTEM MEMORY
(ROM) 131
BIOS 133
(RAM) 132
OPERATING SYSTEM 134
APPLICATION PROGRAMS 135
OTHER PROGRAM MODULES 136
PROGRAM DATA 137
141
PROGRAM DATA 147
OTHER PROGRAM MODULES 146
APPLICATION PROGRAMS 145
OPERATING SYSTEM 144
110
100

FIG. 1

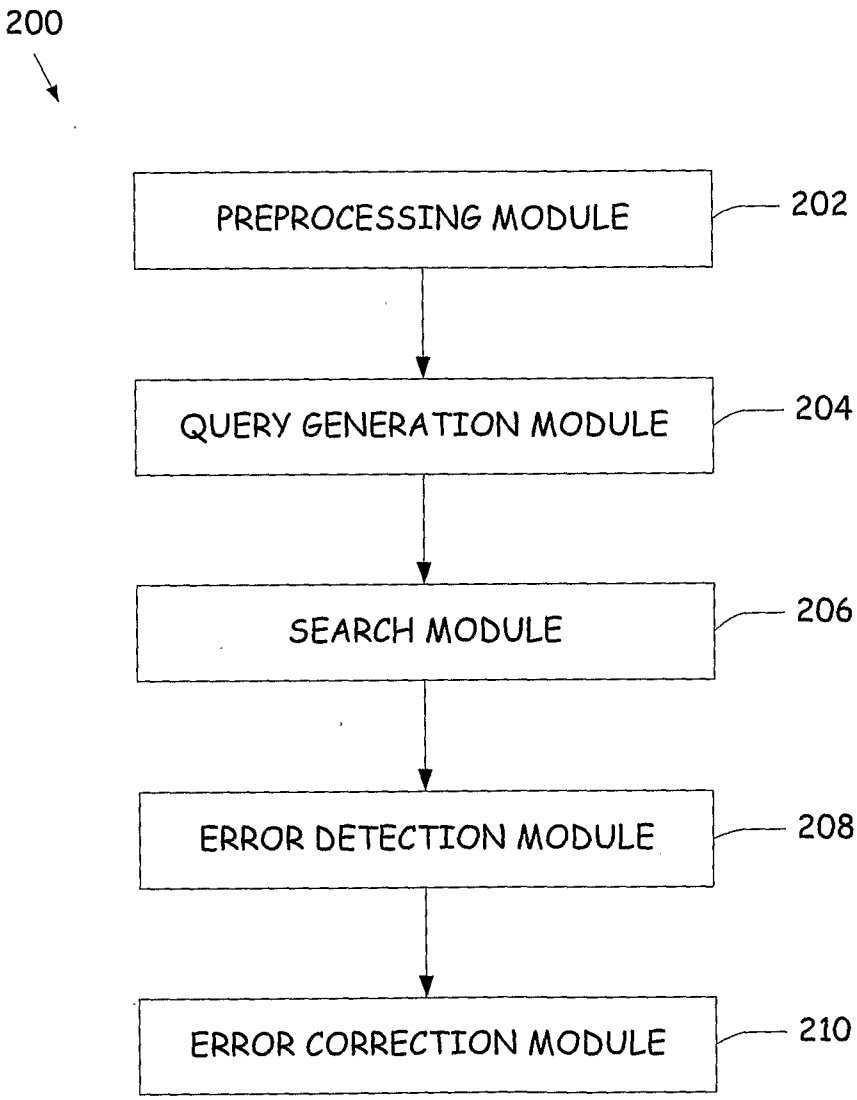

200
PREPROCESSING MODULE
202
QUERY GENERATION MODULE
204
SEARCH MODULE
206
ERROR DETECTION MODULE
208
ERROR CORRECTION MODULE
210

FIG. 2

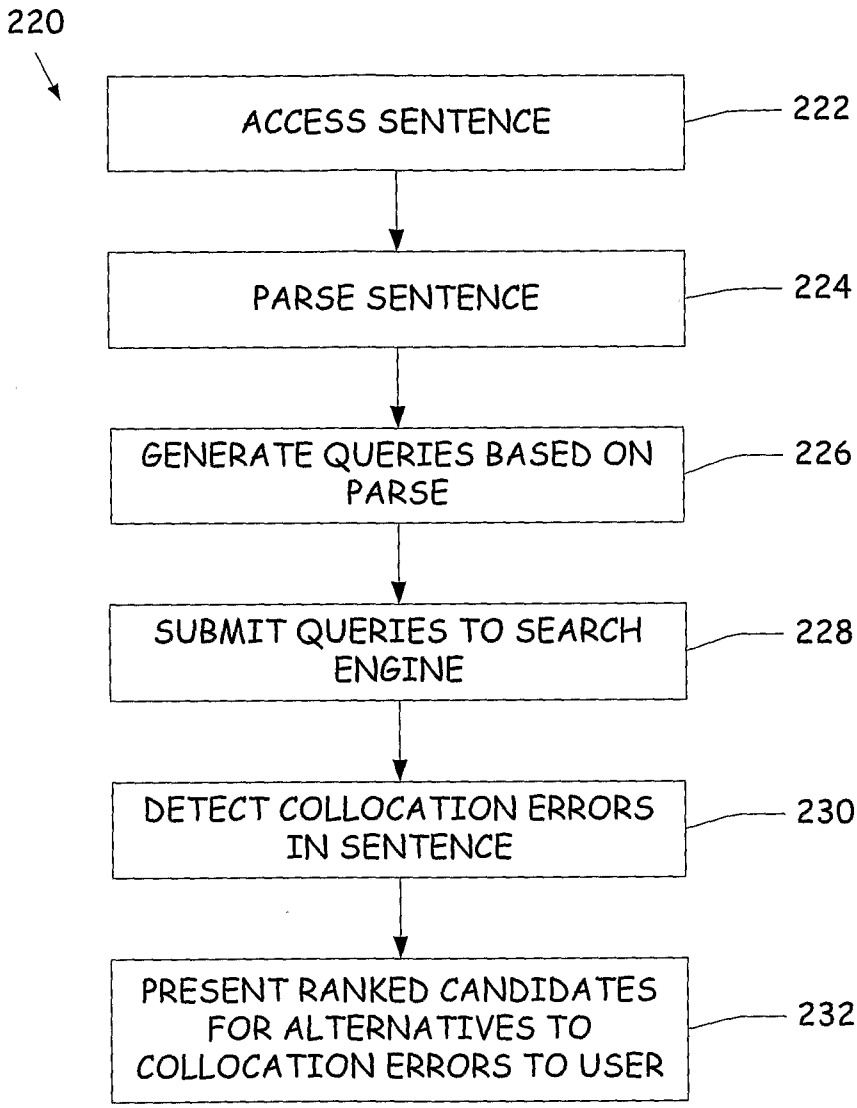

220
ACCESS SENTENCE
222
PARSE SENTENCE
224
GENERATE QUERIES BASED ON PARSE
226
SUBMIT QUERIES TO SEARCH ENGINE
228
DETECT COLLOCATION ERRORS IN SENTENCE
230
PRESENT RANKED CANDIDATES FOR ALTERNATIVES TO COLLOCATION ERRORS TO USER
232

FIG. 3

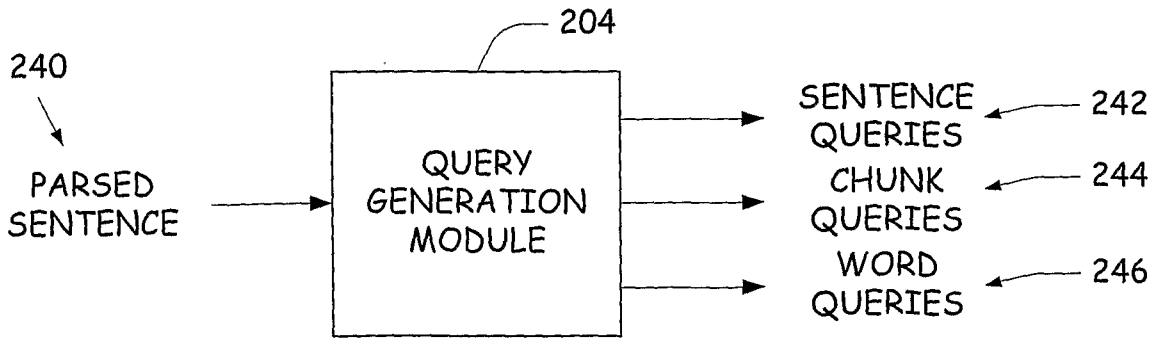

204
240
PARSED SENTENCE
QUERY GENERATION MODULE
SENTENCE QUERIES
242
CHUNK QUERIES
244
WORD QUERIES
246

FIG. 4

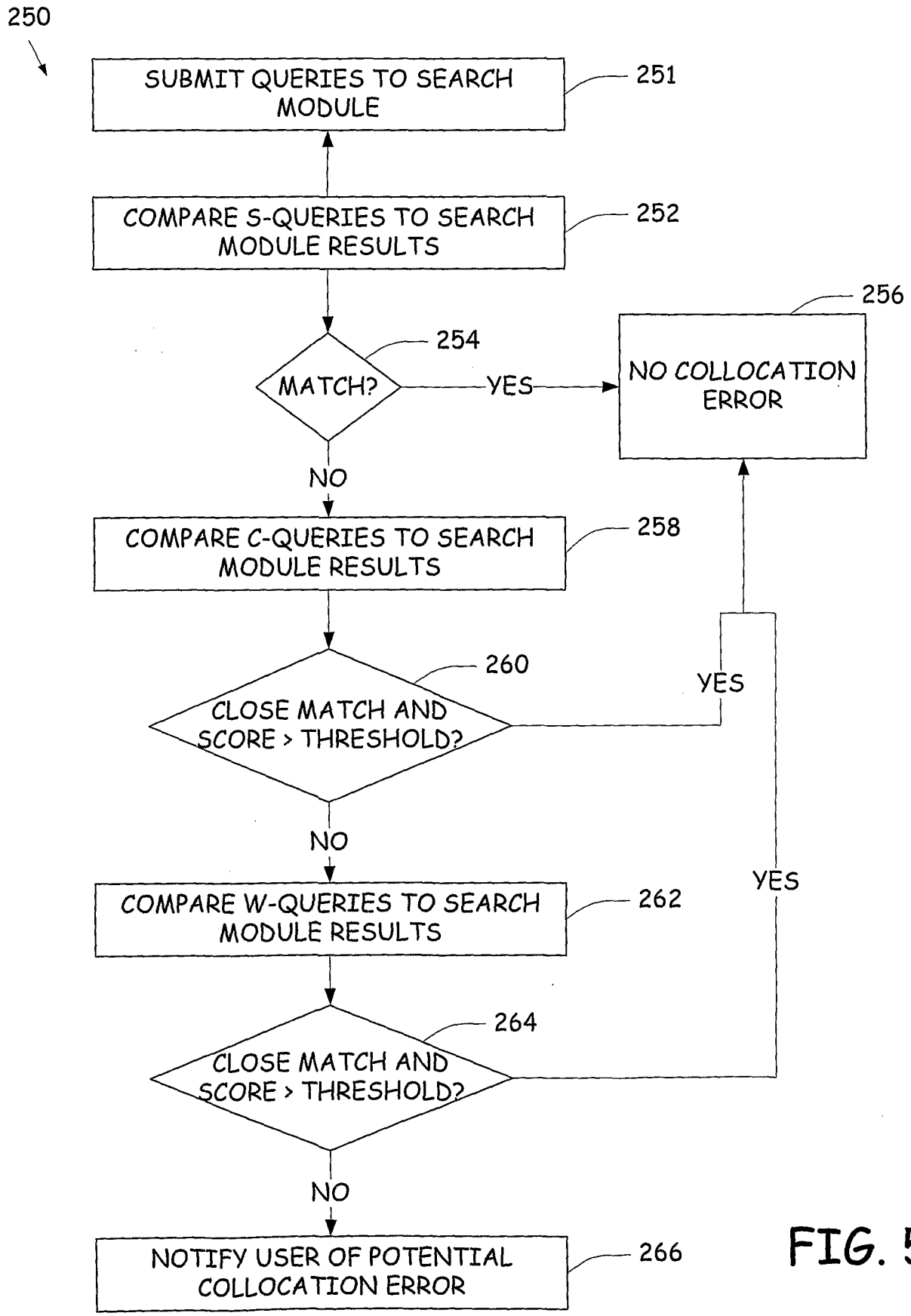

250
SUBMIT QUERIES TO SEARCH MODULE
251
COMPARE S-QUERIES TO SEARCH MODULE RESULTS
252
MATCH?
254
YES
NO COLLOCATION ERROR
256
NO
COMPARE C-QUERIES TO SEARCH MODULE RESULTS
258
CLOSE MATCH AND SCORE > THRESHOLD?
260
YES
NO
COMPARE W-QUERIES TO SEARCH MODULE RESULTS
262
YES
CLOSE MATCH AND SCORE > THRESHOLD?
264
NO
NOTIFY USER OF POTENTIAL COLLOCATION ERROR
266

FIG. 5

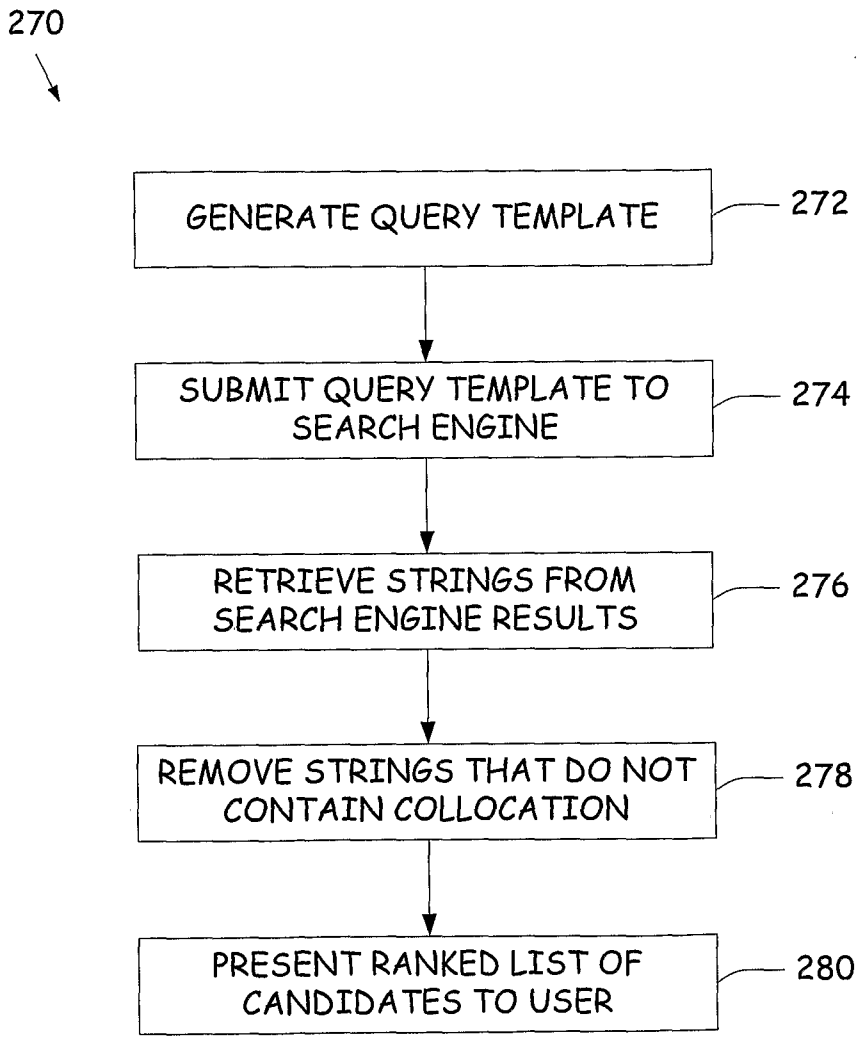

270
GENERATE QUERY TEMPLATE
272
SUBMIT QUERY TEMPLATE TO SEARCH ENGINE
274
RETRIEVE STRINGS FROM SEARCH ENGINE RESULTS
276
REMOVE STRINGS THAT DO NOT CONTAIN COLLOCATION
278
PRESENT RANKED LIST OF CANDIDATES TO USER
280

FIG. 6